Privacy Preserving Data Mining Operations without Disrupting Data Quality

[1]B.Swapna, [2]R.VijayaPrakash

**[1]Department of CSE, SR Engineering College
Warangal, Andhra Pradesh, India**

**[2]Associate Professor in Department of CSE
SR Engineering College
Warangal, Andhra Pradesh, India**

Abstract

Data mining operations help discover business intelligence from historical data. The extracted business intelligence or actionable knowledge helps in taking well informed decisions that leads to profit to the organization that makes use of it. While performing mining privacy of data has to be given utmost importance. To achieve this PPDM (Privacy Preserving Data Mining) came into existence by sanitizing database that prevents discovery of association rules. However, this leads to modification of data and thus disrupting the quality of data. This paper proposes a new technique and algorithms that can perform privacy preserving data mining operations while ensuring that the data quality is not lost. The empirical results revealed that the proposed technique is useful and can be used in real world applications.

Key Words – *data mining, PPDM, sanitization, algorithms*

1. Introduction

Data mining operations have become prevalent as they can extract trends or patterns that help in taking good business decisions. Often they operate on large historical databases or data warehouses to obtain actionable knowledge or business intelligence that helps in taking well informed decisions. In the data mining domain there came many tools to perform data mining operations. These tools are best used to obtain actionable knowledge from data. Manually doing this is not possible as the data is very huge and takes lot of time. Thus the data mining domain is being improved in a rapid pace. While data mining operations are very useful in obtaining business intelligence, they also have some drawbacks that are they get sensitive information from the database. People may misuse the freedom given by obtaining sensitive information illegally. Preserving privacy of data is also important. Towards this end many Privacy Preserving Data Mining (PPDM) algorithms came into existence that sanitize data to prevent data mining algorithms from extracting sensitive information from the databases.

Sensitive information such as the owner of data set or any associated information is illegally taken by data mining operations. This has led to privacy problems in this domain. In order to overcome this there was need for data mining operations that preserve privacy of data. The PPDM algorithms that came into existing to resolvethis problem have given birth to another problem. This problem is the result of sanitization that leads to the disruption of data of decrease in the quality of data. Researches done in [2], [3], [4], and [5] presented the problem of sanitizing data for privacy preserving data mining operations. The reason for the introduction of problems by PPDM algorithms is that these algorithms do not consider two things such as relevance of data and structure of the database into account. With respect torelevance of data, it is possible that all the data present in database is not relevant. This fact has to beconsidered by PPDM algorithms. The structure of dataplays animportant role in data mining operations. The database structure contains table and relationships among them besides having many data integrity and other constraints in place. Data quality is the problem with PPDM operations. However, by considering the database structure and all associated relationships and constraints, it is possible to perform sanitization without disrupting data quality in the underlying database. In order to perform quality-aware sanitization, the PPDM algorithms need some additional information. This is described in other section of this paper.

2. Related Work

The purpose of data mining is to find out new useful patterns or trends or correlations in existing data [7]. This phenomenon can also be called as data pattern processing, data archeology, information harvesting, information discovery, knowledge extraction and knowledge discovery [8]. Database researchers, business communities and MIS communities and statistians use the term data mining. Whereas the term KDD (Knowledge Discovery in Databases) refers to the process of finding useful knowledge from data [8]. Data mining is simply an extension work to the data analysis used for analyzing and understanding the trends in the data. Data mining operations conventionally follow analysis that is a hypothesis driven. It is essentially a domain where patterns and trends are extracted from the historical business data by using certain data mining algorithms. There are two broad approaches for data mining. The first approach is meant for building models that solves problems pertaining to large volumes of data. It is similar to exploratory method and meant for producing summary of data that shows main features of the data [9]. The second approach in data mining is to identify small, sporadic waveforms in EEG traces. For instance unusual spending patterns in credit card usage and so on [10].

Association rule mining is one of the concepts in data mining. There are many algorithms that are used to extract association rules from datasets. The problem with association rules is that the association rules extracted from a data mining operation are more in number that is the cause of concern. The extracted association rules are to be pruned. Pruning of the association rules is an essential operation that is known as post mining operation in the data mining domain [10]. Association rule mining is used traditionally to identify strong association rules in the data. The idea of interesting rule is explored by Agarawal in [11] which allowretrieving only interesting rules out of all association rules. Association rule mining can be done by unauthorized people. To avoid this problem Atallah et al. [12] proposed heuristic techniques that support modification of data with security. Another heuristic based framework for preserving privacy in data mining for association rules is proposed by Oliveira and Zaiane [13]. Their work explores on hiding frequent patterns that contain private or highly sensitive data obtained from the dataset. They achieve it by modifying existing data by inserting noise. The algorithms proposed by them are known as perturbative algorithms. They also propose an approach by name item-restriction approach that allows introducing noise and reducing the removal of real data.

For mining association rules a framework was proposed by Evfimievski et al. [14] which finds association rules from data with categorical values. It randomizes data for preserving the sensitive transactions and still mines true association rules. There is another technique which is also restriction based proposed by Rivzi and Hartsa [15] is known to have a distortion method as preprocess step before actually performing the data mining process with goal to preserve privacy at the individual tulles level for all records.

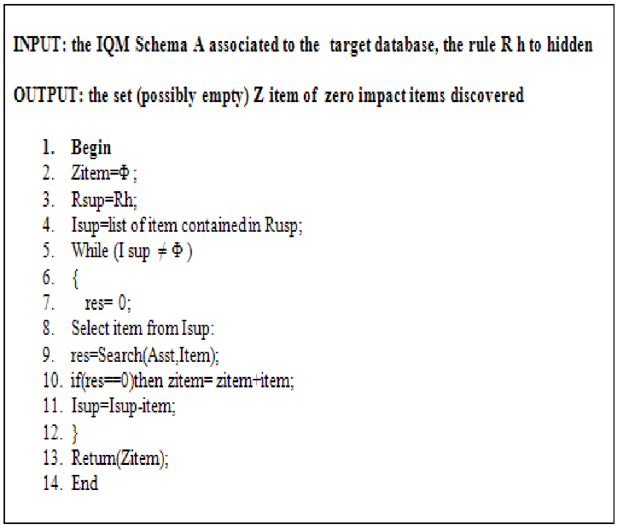
3. Proposed PPDM Sanitization and Data Quality

PPDM algorithms perform data sanitization that prevents accessing association rules from the sanitized database. It does mean that such database allows privacy preserving data mining operations. However the problem identified with such operations is that the quality of underlying database is lost. This paper aims at overcoming this problem by introducing a quality-aware sanitization approach that makes use of additional information pertaining to database before performing sanitization. The data quality refers to the correctness of data between the real world and also the representation of it in the database [6]. The same meaning holds true with respect to PPDM operations too. By measuring real world data and the data which has been sanitized it is possible to assess the quality of data. There are some parameters that can be used to correctly measure the data quality of a database. These parameters [5] include accuracy, completeness and consistency. Accuracy refers to how close the real world data to the data represented in the form of a database. Completeness is a measure to find whether all data items in the real world are being represented by the relational database or not. Consistency refers to the integrity constraints applied to database and how they are true even after performing sanitization.

In this paper we proposed a new information quality model that takes information accuracy, completeness and consistency parameters while making PPDM operations. This model consists of many components. By studying these components it is possible to find whether the given database can be sanitized or not. The very important part of the Information Quality Model (IQM) is the DMG (Data Model Graph) that represents set of attributes involved in the aggregate information including the constraints. Another important aspect is AIS (Aggregate Information Schema) that is used to measure relevancy between different aggregations.

4. Algorithms

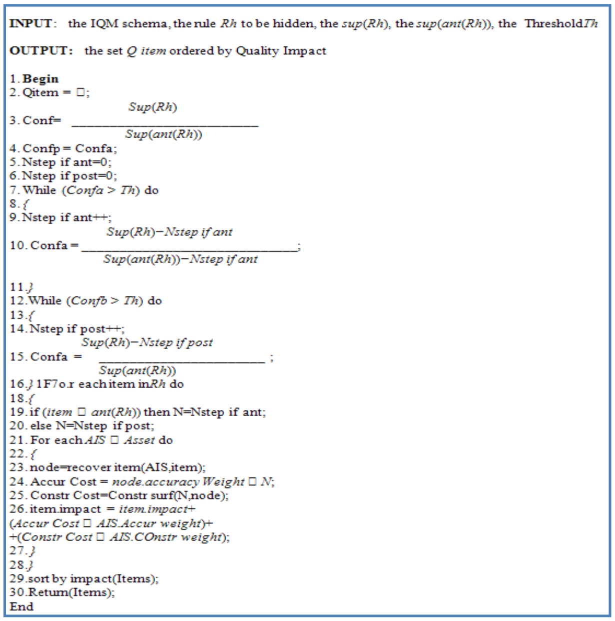
The algorithms proposed in this paper are known as distortion based algorithms. These algorithms are meant for performing sanitization as part of privacy preserving without disrupting quality of database. First of all the rule to be hidden from the association rule extraction algorithms is to be identified using apriori algorithm [1]. Apriori gives set of rules first of all. From the set of rules a sensitive rule that has to be hidden is identified. Next step is to find out zero impact items that are associated with the rule to be hidden. The zero impact algorithmsare given in fig. 1.



```
INPUT: the IQM Schema A associated to the  target database, the rule R h to hidden

OUTPUT: the set (possibly empty) Z item of zero impact items discovered

Begin
Zitem=Φ ;
Rsup=Rh;
Isup=list of item contained in Rusp;
While (I sup ≠ Φ )
{
    res= 0;
Select item from Isup:
res=Search(Asst,Item);
if(res==0)then zitem= zitem+item;
Isup=Isup-item;
}
Return(Zitem);
End
```

Fig. 1 –Zero Impact Algorithm

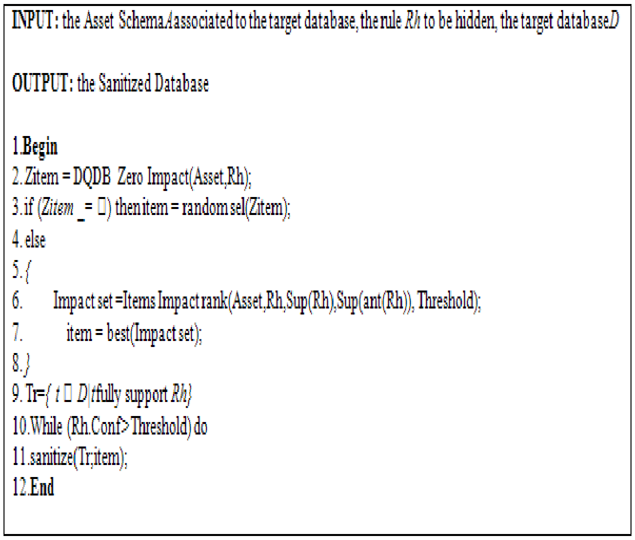
As can be seen in fig. 1 the zero impact algorithm, as the name implies, is used to extract items that have no impact on the quality of data base even when they are modified for the purpose of sanitization. The result of this algorithm is a set of items that have zero impact on the quality of database.



```
INPUT:   the IQM schema, the rule Rh to be hidden, the sup(Rh), the sup(ant(Rh)), the  ThresholdTh

OUTPUT:   the set Q item ordered by Quality Impact

Begin
Qitem = □;
Conf= $\frac{Sup(Rh)}{Sup(ant(Rh))}$
Confp = Confa;
Nstep if ant=0;
Nstep if post=0;
While (Confa > Th) do
{
Nstep if ant++;
Confa = $\frac{Sup(Rh)-Nstep\ if\ ant}{Sup(ant(Rh))-Nstep\ if\ ant}$;

}
While (Confb > Th) do
{
Nstep if post++;
Confa = $\frac{Sup(Rh)-Nstep\ if\ post}{Sup(ant(Rh))}$ ;
} 1F7o.r each item inRh do
{
if (item □ ant(Rh)) then N=Nstep if ant;
else N=Nstep if post;
For each AIS □ Asset do
{
node=recover item(AIS,item);
Accur Cost = node.accuracy Weight □ N;
Constr Cost=Constr surf(N,node);
item.impact = item.impact+
(Accur Cost □ AIS.Accur weight)+
+(Constr Cost □ AIS.COnstr weight);
}
}
sort by impact(Items);
Return(Items);
End
```

Fig. 2 – Item Impact Rank Algorithm

As can be seen in fig. 2 and other algorithms, the sanitization algorithm performs operations described here. All transactions that support the rule to be hidden are selected first. Then zero impact items are computed. Based on that information the rule to be hidden is hidden by modifying the zero impact sets or sanitizing the database.



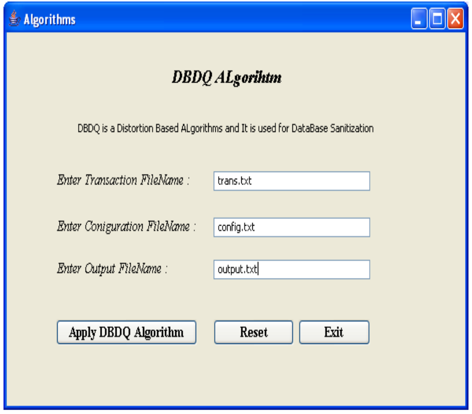
```
INPUT: the Asset SchemaAassociated to the target database, the rule Rh to be hidden, the target databaseD

OUTPUT: the Sanitized Database

Begin
Zitem = DQDB  Zero Impact(Asset,Rh);
if (Zitem _= □) then item = random sel(Zitem);
else
{
    Impact set =Items Impact rank(Asset,Rh,Sup(Rh),Sup(ant(Rh)), Threshold);
      item = best(Impact set);
}
Tr={ t □ D|tfully support Rh}
While (Rh.Conf>Threshold) do
sanitize(Tr;item);
End
```

Fig. 3 – Distortion based sanitization algorithm

As can be seen in fig. 3, the distortion based sanitization algorithm is responsible to sanitize database without reducing the quality of underlying database. This is because this algorithm finds the items that are having zero impact on the quality of database. These kinds of items are used to perform sanitization. The result of this algorithm is the sanitized database which can preserve privacy when PPDM operations are performed on it.

5. Experiments and Results

A prototype application is developed with GUI to demonstrate the usefulness of the proposed algorithm. The environment used in JDK 1.6 (Java Programming Language), Net Beans IDE that runs in a PC containing 2 GB RAM and 2.9x GHz processor. The main screen of the proposed application is as shown in fig. 4.



Algorithms

DBDQ ALgorihtm

DBDQ is a Distortion Based ALgorithms and It is used for DataBase Sanitization

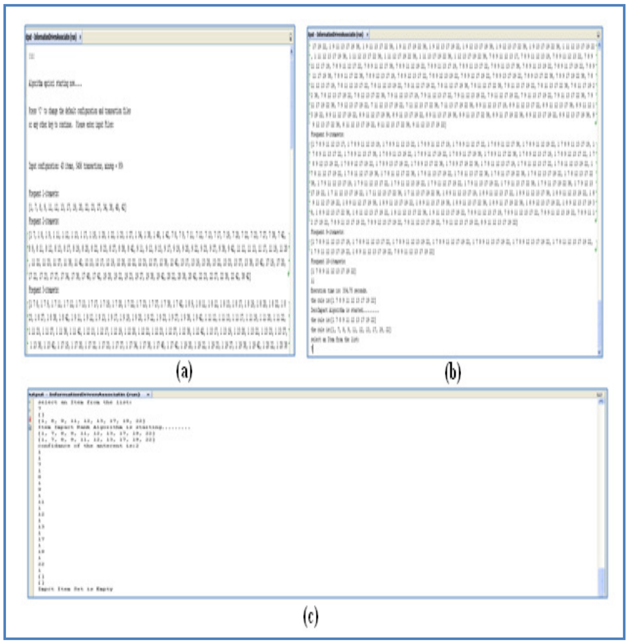
Enter Transaction FileName : trans.txt

Enter Coniguration FileName : config.txt

Enter Output FileName : output.txt

Fig. 4 – The main screen of the application

As can be seen in fig. 4, the application takes one transaction file and one configuration file as input. The output of the algorithm is sent to the given output file. The trans.txt has actual business transactions while the config.txt has required configurations. The DBDQ algorithm is applied on the given data set. The results are shown in fig. 5.



```
select an Item from the list:
Item Impact Rank Algorithm is starting.........
confidence of the antecent is:2
Impact Item Set is Empty
```

Fig. 5 –Experimental results

As can be seen in fig. 5 (a), first of all frequent item sets are generated using apriori algorithm. Then the sensitive rule to be hidden is identified as shown in fig. 5 (b). Afterwards, the hidden rule is used to find zero impact items in the data set. Finally the zero impact item set is shown in fig. 5 (c). If the zero impact items is empty the impact of given item I found and then sanitization is done thus it achieves privacy preserving while eliminating disruption of data.

6. Evaluation

The import of sanitization on data quality, to the best of our knowledge, has not been addressed fully in the past. For this reason the results of this paper could not be compared with other works. However, the data quality aware sanitization technique proposed by us is able to support sanitization without disrupting the quality of the database. Table 1 shows percentage ofvalues for data quality metrics such as accuracy, completeness and consistency.

	Accuracy	Completeness	Consistency
Before Sanitization	100%	100%	100%
After Sanitization(PPDM)	95%	98%	99%
After Sanitization(Proposed)	100%	100%	100%

Table 1 - Data Quality Metrics

The results of the quality metrics are visualized in fig. 6.

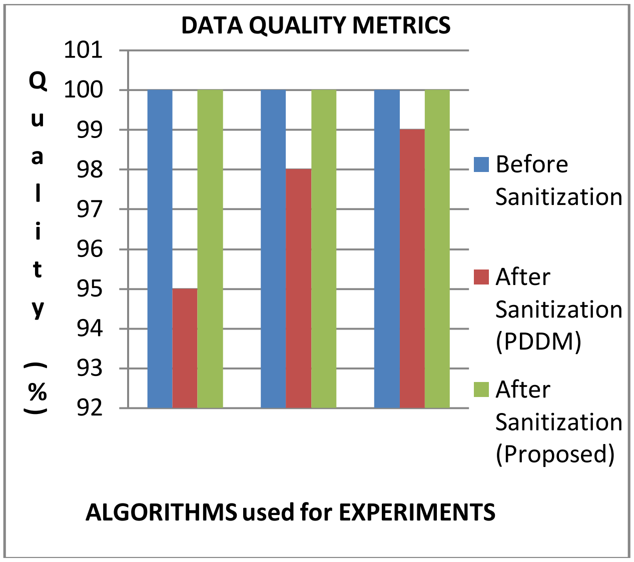


Fig. 6 – Data Quality Metrics

As seen in fig. 6, the data quality metrics after and before sanitization are presented. The metrics used here are accuracy, completeness, and consistency. Accuracy refers to the similarity between the original tuple and sanitized tuple. Completeness t refers to the fact that after sanitization every attribute is not empty. Consistency refers to the fact that integrity constraints remainsatisfied after sanitization.

As shown in fig. 6, a comparison is made among data quality before sanitization, aftersanitization with existing PPDM algorithms and the proposedalgorithms in this paper. The results reveal that the proposed algorithms in this paper are capable of performing sanitization without disrupting the data quality.

7. Conclusion

The PPDM algorithms that came into existence were achieving data privacy while performing data mining operations. However, these algorithms achieved this by sanitizing datathus leading to the reduction of data quality. In order to overcome this drawbackthis paper proposes a new technique and algorithms that are used to allow privacy preserving data mining operations without disrupting the quality of underlying database. We have developed a prototype application that demonstrates the efficiency of the proposed algorithms for privacy preserving data mining while keeping data quality intact.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *In Proceeding of the 20*th *International Conference on Very Large Databases, Santiago, Chile, Morgan Kaufmann*, June 1994.

[2] I.M. Author. A framework for evaluating privacy preserving data mining algorithms*. *Journal Data Mining and Knowledge Discovery*, 11(2):121–154, September 1999.

[3] E.Bertino and I. Fovino. Information driven evaluation of data hiding algorithms. In 7th *International Conference on Data Warehousing and Knowledge Discovery*. Springer- Verlag, 2005.

[4] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–228, New York, NY, USA, 2002. ACM.

[5] U. ofMilan Computer Technology Institute Sabanci University. Codmine ist project. 2002-2003.

[6] K. Orr. Data quality and systems theory. *Commun. ACM*, 41(2):66–71, 1998.

[7] Chung, H. M., Gray, P. (1999), "Special Section: Data Mining". Journal of Management Information Systems, (16:1),11-17.

[8] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, R (1996). "The KDD Process for
Extracting Useful Knowledge from Volumes of Data," Communications of the ACM, (39:11), pp. 27-34.

[9] Hand, J., Kamber, M. (2001), Data Mining: Concepts and Techniques, Morgan-Kaufmann Academic Press, San Francisco. Hand, D. J. (1998), "Data Mining: Statistics and More?", The American Statistician, May (52:2), 112-118.

[10] Rajagopalan, B., Krovi, R. (2002), "Benchmarking Data Mining Algorithms", Journal of Database Management, Jan-Mar, 13, 25-36

[11] Rochlani, Yogesh R., and A. R. Itkikar. "Integrating Heterogeneous Data Sources Using XML Mediator." ijcsn, vol 1, issue 3.

[12] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure limitation of sensitive rules. In Proceedings of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pages 45–52. IEEE, 1999.

[13] S. R. M. Oliveira and O. R. Za ýane. Privacy preserving frequent itemset mining. In CRPIT '14: Proceedings of the IEEE international conference on Privacy, security and data mining, pages 43–54, Darlinghurst, Australia, Australia, 2002. Australian Computer Society, Inc.

[14] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 217–228, New York, NY, USA, 2002. ACM.

[15] S. J. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases, pages 682–693. VLDB Endowment, 2002.

[16] R. Srikant and R. Agrawal. Mining generalized association rules. In Proceedings of the 21th International Conference on Very Large Data Bases, pages 407–419. Morgan Kaufmann, 1995.